

DOCUMENT RESUME

ED 480 079

CG 032 652

AUTHOR Harris, William G.
TITLE Current Issues in Educational Assessment: The Test Publisher's Role.
PUB DATE 2003-08-00
NOTE 26p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE Information Analyses (070)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS *Accountability; *Educational Assessment; Educational Testing; Elementary Secondary Education; *High Stakes Tests; Public Education; *Test Construction; Test Use
IDENTIFIERS *Test Publishers

ABSTRACT

As education policymakers have moved to reform K-12 public education, the roles of test publishers in assessment have expanded. In the last two decades these expanded roles have coincided with the movement of assessment to the center of education reform initiatives. The drive for improvement in public education has made the roles of test publishers even more demanding while presenting the publishers with new opportunities and challenges. This chapter reviews the multifaceted role of educational test publishers, as well as the demands place on standardized assessments and assessments used in high-stakes decisions. (Contains 34 references.) (GCP)

Current Issues in Educational Assessment: The Test Publisher's Role

By
William G. Harris

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- (1) This document has been reproduced as received from the person or organization originating it.
- (2) Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



Chapter 44

Current Issues in Educational Assessment

The Test Publisher's Role

William G. Harris

As education policymakers have moved to reform K–12 public education, the roles of test publishers in assessment have expanded. In the last two decades these expanded roles have coincided with the movement of assessment to the center of education reform initiatives. In the 1980s, users of assessments largely focused on minimal competency testing. By the 1990s, education policymakers had ratcheted up the expectations. The focus changed to high-stakes accountability in which the assessment served as the leading indicator and, unfortunately, in some instances as the only indicator (Linn, 2000). The drive for improvement in public education has made the roles of test publishers even more demanding while presenting the publishers with new opportunities and challenges.¹

An educational assessment is a standardized method of gathering data and converting it to information used to evaluate the academic progress of students, the effectiveness of instruction, or the success of educational programs (Cizek, 1997). Ideally, most jurisdictions employ multiple measures for each purpose—such as standardized tests, writing samples, portfolio materials, and teachers' recommendations—to create an educational assessment system for measuring different elements of academic achievement or for evaluating a state or district's overall program performance. For the purpose of this discussion, I define educational assessment specifically as (a) standardized testing used by teachers to identify strengths and weaknesses of students in order to adjust classroom instruction; (b) standardized testing used in making high-stakes decisions such as grade promotion and graduation; or (c) the aggregation of non-student-specific standardized testing data used to make program decisions such as educational funding and school staffing. It is extremely important to identify the type of standardized testing at issue so that a proper context for discussion is available.

Most stakeholders such as education policymakers, educators, and parents embrace the importance of assessment in educational or instructional improvement. Such widespread support begins to waver, however, when the assessments possess high-stakes consequences, which morphs the test into a feature of educational policy. Differences among stakeholders surface on the frequency of testing, its overall weight in academic and programmatic accountability, and its influence on the funding of educational resources. The role of the publishers of all types of tests is first to recognize the legitimacy of the differences and then to campaign energetically for the appropriate and meaningful use of all assessments in an education reform strategy.

Assessments used for high-stakes purposes serve as the gatekeepers of the standards-based accountability reform movement.² Standards-based reform refers to the use of state standards for subject matter content (such as mathematics, language arts, or other core subjects in each grade) and to the use of performance levels established by the state for determining if students are performing at acceptable levels of competency (such as "Basic," "Proficient," or "Advanced"). Accountability means that parents, students, educators, and policymakers share the responsibility for improving the academic achievement of students in accordance with specific content and performance standards. Educational assessments are central to the standards-based reform system that stresses the use of measurable outcomes to monitor students' progress. In states that have implemented graduation assessments, however, adverse reactions of parents, teachers, and educators, as well as uncertainty among policymakers, have led to extensions or delays in imposing those graduation requirements.

On top of the academic results, most states and districts have implemented an accountability system for measuring programmatic progress. Some states have even adopted systems for rewarding or sanctioning schools or districts based upon those outcomes. Because of the uncertainties surrounding these accountability measures, many policymakers have delayed implementation of specific rewards or sanctions.³

The more that stakeholders depend on educational assessments to direct policy, the more test publishers are placed in the role of securing validity evidence to support high-stakes uses while discouraging the use of any one assessment as a sole determinant in these decisions. Generally accepted professional technical standards emphasize the use of multiple measures especially when the assessment outcomes are tied to high-stakes consequences. In that scenario, test publishers emphasize

the value of educational assessments but point to the importance of multiple measures to provide complementary or confirmatory information to aid in the decision-making effort.

A Multifaceted Role

At a strategic level, the roles of educational test publishers are not easily partitioned into discrete functions. The interrelatedness of various roles points to a single role that is multifaceted in its composition. The strategic objectives inherent in the test publishers' multifaceted role are compatible across stakeholder groups. A test publisher's materials may convey the concept of test validity and test fairness differently to education policymakers, educators, and parents. The intent is to assure each of these groups that the inferences drawn from an educational assessment are accurate and that the assessment outcomes do not lead to uneven or unfair treatment of students. Success in managing the test publisher's multifaceted role depends on effective communication of the way a particular assessment functions in the accountability system. As such, the test publisher is strategically compelled to communicate the right information at the right level of understanding to the right stakeholder (e.g., students, parents, educators, policymakers).

A test publisher's multifaceted role is largely molded out of a business necessity, yet this situation creates values and benefits that extend well beyond mere business interests. For instance, a well-designed, professionally developed educational assessment can contribute to understanding the alignment between state content standards and curriculum, to improving the quality of educational diagnostics, to targeting the educational resource needs of low-performing schools, and to monitoring efforts to afford all students the opportunity to learn. When psychometrically supported and appropriately used, the educational assessment adds value to an educational improvement strategy and contributes, both socially and educationally, to the greater good of society.

The broad influence of educational assessments creates for test publishers both opportunities and challenges. As already suggested, some of the opportunities are in educational diagnostics, decision making (e.g., graduation and promotion examinations), classroom instruction, and intervention or remediation strategies. Safeguarding educational assessments from misuse, unreasonable criticism, and misperceptions are among the challenges test publishers face. Another

equally important challenge is anticipating and planning for the interplay between assessments and technology.

In its multifaceted role, a test publisher attempts to communicate the appropriate function of assessment in the educational process. The test publisher circumscribes the capabilities of a specific educational assessment as effective when its purpose is well defined and its use does not stray from its intended purpose. Several issues ruffle the neatness of this statement. A particular educational assessment may generate useful information about the performance of an individual student, a group of students, or an educational program. The same assessment may be valid for more than one purpose and in multiple settings. As such, there may be a wide range of appropriate use of some assessments.

Despite stakeholders' heavy reliance on educational assessments, however, assessments are incapable of closing the achievement gap between students from high-performing schools and those from low-performing schools. Assessments offer policymakers and educators guidance on ways to close the gap, but they, as part of standards-based accountability reforms, are powerless to correct long-standing problems of educational indifference. Therefore, it is untenable to burden educational assessments with the task of improving the quality of education without policymakers aggressively addressing factors such as inadequate per-pupil expenditures, unacceptable pupil-teacher ratios, and ill-equipped classroom teachers. When these and related factors (e.g., educational intervention at the prekindergarten level) are addressed with a sustained commitment, the benefits of educational assessment are attainable.

Put differently, a classroom environment that is resource starved and pedagogically shortsighted undermines both learning and the benefits of the educational assessment. Narrowly "teaching to the test" strips the assessment of its value and shortchanges the education of students. On the other hand, when inadequacies in the classroom environment are corrected in concert with the use of a professionally developed assessment, students are given the chance to become better learners, rather than merely better test takers.

In their communicator role, test publishers seek to explain that an accountability system of content and performance standards and assessment is inadequate to sustain long-lasting, meaningful reform. The absence of real changes in the classroom environment, in teacher development, and in technology use marginalizes both the standards and assessment in schools with students who could benefit the most

from them. Such tension, if not properly addressed, can only accelerate the erosion of confidence in the reform effort and, perhaps, in the specific educational assessment selected for use in particular states or districts.

A key skill for the test publisher, then, is to perfect the ability to find the appropriate level at which to communicate relevant information to different stakeholder groups. For instance, it is vitally important to explain to teachers the disservice they provide to students when they teach to the test. Such inappropriate test preparation hampers true learning and potentially discolors the usefulness of the test results. Clear, thoughtful, and realistic content standards that encourage the development of a rich, vibrant curriculum are pivotal to any effort to avoid turning the classroom into a test prep shop. As a communicator, the test publisher campaigns continually for stakeholders to use sound testing practices and to integrate the educational assessment into the learning experience of students. As the assessment becomes integrated in learning, it is less likely to be the target of disillusioned stakeholders and testing critics.

Reforming Education and the Educational Assessment

As noted, through legislative reform initiatives that emphasize standards-based accountability, policymakers and educators have fueled the growth of the educational assessment. Such growth has assigned to test publishers a position of influence in the movement to reform the nation's K-12 public education system. The influential role of test publishers and the spiraling rise in testing are events that have evolved over the past two decades.

By the early 1980s, policymakers and educators had sounded the alarm that the nation's education system was performing poorly and that the whole system required a radical overhaul. They assailed the nation's education system as inefficient and ineffective. The inadequacies of a burdened education system produced students of low academic achievement.

In decrying the plight of the education system, policymakers and educators were not alone. Business leaders added their voices to the chorus of critics urging the reinvention of public education. These leaders linked a quality education to the country's future economic security and global competitiveness. They offered mostly anecdotal evidence to support their claims that without a vibrant education system, the business prowess of the United States would suffer increased threats. Such threats from competitive forces were expected to intensify because

the nation's education system was fractionated and ill equipped to prepare students to join a technologically demanding workforce. Businesses lamented that often they were forced to provide remedial education to high school graduates or look outside the United States to find employees with the prerequisite skills, training, and education. For these leaders a quality education had become a business imperative.

Despite these needs, meaningful comparisons of student achievement across the 50 states proved elusive. The problem in comparing the 50 state education systems existed in part because each state employed different educational assessment instruments and different testing cycles for different grade levels. With education as primarily the dominion of the state, attempts to equate different commercially published instruments used by states met with only meager success, except for limited situations, such as for assessments used to measure progress among impoverished children. Adding to this complexity was significant state variation in the level of educational expenditures, curriculum content, and standards for measuring student achievement. Cross-state comparisons were fraught with methodological pitfalls, and comparisons of students within the same state were not without limitations due to the use of different local tests by various districts across a state. Even with these methodological barriers, the use of nationally normed, standardized large-scale tests was the best available alternative for measuring the student progress and the success of educational programs.

In 1983, concerns about the nation's education system were confirmed with the release of the National Commission on Excellence in Education's final report, *A Nation at Risk*. That report acknowledged and highlighted deep systemic problems in the nation's education system. It pointed out that the content of school curricula and measurable standards of accountability were woefully inept and needed to be upgraded. The report also called for students to devote more time to learning and for teachers to receive more resources to improve teaching preparation. Although the report has had its critics, it has served, albeit with changes, as a national blueprint for the standards-based education reform movement.

By the 1990s, both a Republican and a Democratic president had reacted to that report by seeking legislation to encourage states to improve their standards-based reform efforts. Initially, after President George H. Bush's education summit of governors and business representatives recommended a series of National Education Goals, he introduced the America 2000 legislation to provide federal money for

states to engage in systemic education reform focused on standards and assessments. Picked up, revised, and renamed by President Clinton as Goals 2000, the legislation was enacted into law in 1994 as the Educate America Act with the avowed aim of having states adopt "world-class content standards and break-the-mold assessments to measure them"⁴ (p. 8) By 1996, every state had accepted federal funds for these purposes, and to date, nearly every state has developed its own set of content standards; 47 states have adopted some form of assessment system to measure that content.

Criticizing Education Reform and Assessment

The assessment component of the education reform movement has received a disproportionate amount of attention and criticism. Assessment represents only one of the key activities of education reform. Education reform contains two major branches of activities: resource allocation and structural reforms (Grissmer, Flanagan, Kawata, & Williamson, 2000). Resource allocation reforms target factors such as per-pupil expenditures, teachers' salaries, pupil-teacher ratios, and teachers' resources. Structural or standards-based reforms target the development of well-designed, realistic content standards aligned to state curricula, which can then be used to develop assessments. Educational assessments are used to measure directly the effects of standards-based curriculum and to measure indirectly the effects of resource allocations on student achievement and educational programs.

As the standards-based reform movement has charged forward, its reliance on assessment has provoked criticism. The level of resistance to assessment varies among proponents and opponents of reforms. Some proponents of education reforms complain that a standards-based accountability system prematurely places too much emphasis on testing with high-stakes implications. They view the tendency "to rush to test" as outpacing a balanced approach to education reform. Yet there is little disagreement that assessment is fundamental to an effective standards-based accountability system; it seems that testing creates the most concern when it is first introduced. The introduction of large-scale standardized testing is meant to improve education and instruction, not distract from it. This desired use encourages teachers and educators to redesign the curriculum, to establish teacher preparation programs, and to create intervention and remediation programs that reflect clearly defined content standards. These activities are not high stakes because they are not used to make individual student decisions. For state testing

proponents, the key drivers are content standards. State assessments used for these purposes provide the classroom teacher, as well as each student's parents, with specific information on student strengths and weaknesses in particular subjects within the state's content standards. Such state standardized assessments have been developed to stimulate a productive learning environment rather than one regimented around test preparation.

In most settings where these state standardized tests are used, except where high school graduation itself is the purpose,⁵ many other factors exist from which individual decisions about student placement and promotion are made: grades, portfolios or simple writing samples, teacher recommendations, attendance, extracurricular activities, and the like. It is not appropriate or fair to label these tests as automatically having a high-stakes purpose when the most common use of information is directly by teachers and educators, to guide classroom instruction and intervention or remediation for students.

Using these tests to provide program information is also not a problem. In the most common situation, districts or states will take the aggregated data from their standardized tests, without any identifiable student information, and disaggregate the data. In other words, states and districts are able to determine based on general data how specific subgroups of students (e.g., by race, ethnicity, gender, type of disability, or family income level) are performing against the state content standards. These disaggregated data are used to determine whether the subgroups are "narrowing the gap" with all other students.

Evaluating the Criticism

Some critics insist that too much instructional time and curriculum content is lost to test preparation and test taking. They argue that students are shortchanged because extracurricular activities such as music and art vanish from the curriculum and are replaced with a concentrated effort to teach to the test. They further assert that the growing obsession with accountability and test results narrows the curriculum and stymies creativity. Still, there is nothing intrinsically limiting about using state assessments for instructional purposes.

Other critics assert that the opportunity to learn is grossly uneven for students from low-performing schools and that state standardized assessments further injure them. Students in these schools produce predictably lower scores and their scores are then used to imply that they are less capable than students from high-performing schools. These

critics contend that scores on state standardized assessments for students from low-performing schools are difficult to interpret because the gap in instructional resources rivals the gap in achievement scores for high- and low-performing schools. The subpar test scores of students trapped in marginal schools merely subtract from their already low self-esteem. Subsequently, critics are quick to question the instructional purpose of educational assessments. As a remedy they urge greater emphasis on interventions to provide students with greater opportunities to learn (e.g., better facilities, better prepared teachers, smaller class sizes, instructional resources) and less emphasis, at least initially, on test scores. In responding to these critics it is clear that low-performing students stand to gain the most from assessments when teachers use test results to develop and employ strong intervention and remediation strategies. Shortcomings stem from the failure of the state or locality to provide adequate resources, not the use of valid assessments.

Popham (1999a, 2001) insists that typical state standardized assessments are both misnamed and misleading. He opposes the makeup of traditional assessments while embracing the educational assessment engineered to fit his model. Popham (1999b) views state assessments as overly focused on accountability issues and argues that the assessment of instruction is absent in the test design used to construct these state assessments. In the short run, Popham recommends avoiding the use of these assessments to appraise instruction. He offers an all or nothing perspective on existing educational assessment programs. It is unreasonable to ignore the instructional benefits derived from existing state standardized tests. Nevertheless, Popham's recommendation to design state tests capable of measuring both instruction and overall accountability is compelling and is a potentially beneficial refinement.

At another level, Popham (1999a) criticizes state assessments for their inclusion of too many items that measure what students bring to school and not what they learn there. Students from affluent schools come to school with rich and varied life experiences that are captured in the content of many standardized assessment items (Popham, 1999b, 2001). In an attempt to advance his perspective, CISA (2001a) has codified Popham's recommendations in a model RFP with nine requirements for states to design tests that promote better teaching and learning. Five leading education groups, including a panel of prominent educators and measurement specialists, endorse this model RFP (CISA, 2001b). Popham's contention that state-specific items developed in conjunction with state educators and teachers are poorly constructed is not well documented. Items developed without regard to measurement

principles usually reveal substandard psychometric properties. This is rarely the case for state assessment items. Current standardized state assessments are objective measures of state content standards, which are based on professional norms and psychometric rigor.

The heightened position of assessments in education reform leads to sharpened criticism and intensified calls for alternatives. Testing critics serve as a source of information about the function of assessments in education reform. Publishers are seldom in a position to ignore criticism of testing; instead they try to incorporate criticism, when feasible, into an ongoing test improvement strategy.

Advocating for the Educational Assessment

In advocating the indispensable role of the educational assessment in public education, the test publisher also champions its social value. At one level embracing the social value of high-quality education reform is strategically consistent with business objectives. At another level expressing the social value of the educational assessment and educational improvement is a social responsibility. When the assessment truly meets the demands of the education community and society at large, the business objectives of the test publishers are invariably met.

An educational assessment properly aligned to state standards and the curriculum reveals more than the academic progress of students. The assessment discloses how well and how evenly education reforms are serving all students. The newest federal initiative, NCLB, requires more than the regular assessment of students.⁶ Assessment is part of the frontline effort to revamp an education system tattered and frayed in certain respects by providing both longitudinal and cross-sectional data about student progress using each state's own test system. NCLB requires a confirmation by which the state's tests can be generally evaluated. Finally, state measures of "adequate yearly progress" will be reviewed by the U.S. Department of Education and, where appropriate, intervention strategies will be implemented for districts or schools that are not meeting academic improvement expectations.

A wave of recent surveys on educational issues reveals that stakeholders, including parents, describe education in low-income schools as in crisis. These respondents are far less inclined to assign a similar description to high- or middle-income schools (Hart & Teeter, 2001). Schools in low-income areas struggle with overcrowded classrooms, outdated textbooks, ineffective remediation services, too few highly trained teachers, and a host of related school resource issues.

To withdraw the standardized assessment from students in these educationally needy schools would be misguided as well as a disservice to the core meaning of education reform for all students.

All students, teachers, and school administrators need to know how well they measure up to well-defined standards. The social value of professionally developed assessments is in contributing to an intervention and remediation plan that is comprehensive and inclusive. Such a plan does not minimize strong accountability standards or shortchange instruction. Converting score information to a relevant, clearly defined plan for students and programs is the hallmark of a responsible accountability program. To expect anything less from standardized assessments is to emphasize scores at the expense of real reform and an improved educational experience for all students.

The failure to translate state assessment results into educational solutions invites resistance to standardized assessments. To put it more succinctly, generating assessment results without a clear purpose is a misuse of that assessment. The resistance to such misguided actions emerges as complaints of too much testing, boycotts, or initiatives to reduce the influence of the assessment on education reform. Surprisingly, complaints and boycotts of the assessment are less likely to come from stakeholders whose constituents are represented in the low-performing schools. These parents accept, however reluctantly, that the potential benefits derived from the educational assessment outweigh their concerns. Parents in high- and middle-income school areas are more likely to voice discontent about state-mandated content standards, large-scale state assessments, and their supposedly stifling effect on school curriculum.

Recent boycotts and protests of educational assessments in the states of New York, Massachusetts, Arizona, and Illinois further illustrate some parents' growing dissatisfaction (Zernike, 2001). These parents strongly support high standards and demand that their children perform at the higher end of the achievement continuum. They do not, however, endorse standardized assessment as the best way to measure the quality of education. "These kinds of tests reduce content, they reduce imagination, they limit complex curriculum, they add stress and cost money," explains one Scarsdale, New York, parent (Hartocollis, 2001, p. D2). This tremor of discontent is troubling. More importantly, it serves as a signal to test publishers that the success of students on a state assessment does not always equate to unwavering support for testing. Parents contend that state assessments limit the curriculum, curb the use of innovative teaching methods, and suppress creative

thinking among students. These are examples of criticism that test publishers need to address. Finding ways to fashion such discontent into benefits of educational assessment adds value to students' academic experiences and increases parental support for large-scale assessment programs.

Besides parents fearing that state assessments adversely affect creativity and learning, there are other reasons stakeholders retreat from assessment. This withdrawal occurs when the assessment is misaligned with the standards and curriculum, and is then improperly linked to high-stakes consequences, such as graduation. In this situation, unreasonably high standards that focus on extremely high performance levels or that are outside the curriculum actually being taught are allowed to shape the development of the state assessment. This scenario illustrates that, even if the content standards and the state assessment are aligned, if actual curriculum and teaching are not tied to the content standards for the result can be disastrous. Because the state test does not fit the educational reality of what teachers are teaching and students are learning, poor test outcomes occur, which inflame students, parents, and educators. The proclivity of disgruntled parents, educators, and in some cases, the media, is to attack the state assessment as inaccurate and poorly designed. Often these stakeholders call for a moratorium on the use of the assessment for high-stakes decisions. Such misalignment problems are generally discovered during the pretesting phase of developing the assessment instrument. Still, test publishers cannot be perceived as providing merely a "plug and play" assessment device without accepting a growing threat from some stakeholders to reduce the involvement of high-stakes assessment in education reform.

Advocating for the importance of standardized assessment is inseparable from the broader activity of advocating for a quality education. A professionally developed assessment instrument is unlikely to survive untarnished in an education system where the other components are not constructed with the same meticulous care. As an advocate, the test publisher's responsibility does not begin and end with the educational assessment. The responsibility of the test publisher extends to proposing refinements to standards, providing insight into ways to create multiple measures that truly complement the assessment, and finding ways to fold salient concerns of parents and teachers into the assessment effort.

Safeguarding Educational Assessments from Threats

The installation of tough standards-based accountability systems with high-stakes assessments as the linchpin of reform holds some risk for test publishers. Testing with high-stakes consequences puts pressure on test validity, security, and other elements of technical quality (Carnevale & Kimmel, 1997). This pressure increases when education policymakers stretch the test purpose beyond its normal limits. For example, the use of test scores to decide bonuses for teachers generally stretches the test beyond its intended purpose. Using test scores alone represents a misuse of the test; administrators have available other factors to use in conjunction with student test scores, including evaluations by supervisors or the principal, review of lesson plans, parent complaints and accolades, teacher attendance, training records, and the like.

The misuses of large-scale standardized high-stakes assessments were a driving force that led the U.S. Department of Education Office for Civil Rights to develop a guide for policymakers and educators entitled *The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resource Guide for Educators and Policy-Makers* (OCR, 2000). The *Resource Guide* informs policymakers and educators about the interplay among large-scale assessments, professional technical test development principles, and federal nondiscrimination laws. The overarching principles of the *Resource Guide* are culled from a report prepared by the National Research Council entitled *High Stakes: Testing for Tracking, Promotion, and Graduation* (Heubert & Hauser, 1999). These principles are that (a) a test be valid for a particular purpose; (b2) a test reflect the knowledge and skills covered in instruction; and (c) scores on a test lead to decisions and to intended and unintended consequences that are educationally beneficial. As this report makes abundantly clear, when stakeholders employ an assessment as the locus of decision making, it is important that they not unwittingly gloss over the implications of the test or the practices that surround its use or misuse.

Some test practices, when compared against the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and the *Code of Fair Testing Practices in Education* (JCTP, 2002), fall short of these generally accepted professional principles. Occasionally, test practices fall short of existing federal constitutional, statutory, and regulatory nondiscrimination principles. These legal principles address assessment issues such as (a) test use that is incompatible with test

design and validity evidence; (b) the use of a test score as a sole determinant for making decisions; (c) the opportunity for students to receive quality classroom instruction before taking a high-stakes assessment; (d) the significance of fairness being evident in the assessment system; and (e) the educational rationale for establishing cutoff scores. Legal principles are invoked whenever improper use of the educational assessment is alleged in one of these areas.⁷

Although the analysis of relevant federal court decisions cannot be pursued in this chapter, most of the issues confronting the courts regarding the use of educational assessments for high-stakes purposes are directly relevant to test publishers. The more the assessment results disproportionately affect the educational experience and success of certain groups of students (e.g., minority groups, students with limited English proficiency, or students with disabilities), the more probable the assessment will be embroiled in litigation.⁸ The *High Stakes* report (Heubert & Hauser, 1999) stopped short of calling for federal regulation of high-stakes assessments, but it does argue that the two major mechanisms for compelling appropriate test use—voluntary compliance with professional technical standards, such as the *Standards* (AERA, APA, & NCME, 1999), and legal actions—are inadequate. This call for tighter control of the assessment process echoes from groups such as the National Commission on Testing and Public Policy (1990) and preparatory organizations (Katzman & Hodas, 2001).

The *OCR Resource Guide*, more than any other recent document dealing with testing issues, serves as a bridge between the *Standards* and relevant legal standards. It offers practical guidance to stakeholders on appropriate use of assessments for high-stakes decisions and on the legal pitfalls to eschew when using these assessments in accountability systems. Relying on the *Resource Guide* as part of an aggressive preventive outreach program would diminish markedly the need to entertain regulatory remedies for inappropriate test use. Test publishers continue to advocate the benefits of the *Resource Guide*, and have urged the Department of Education to create a substantial outreach program for all stakeholders.

Besides ensuring proper use of large-scale state assessments used in high-stakes decisions, it is important to safeguard their integrity. One of the most common threats to the integrity of assessments is cheating. In May 2001, several Maryland teachers used the actual state sixth-grade mathematics test as practice for their students. Ironically, it was the students themselves who blew the whistle by telling other teachers they had seen the items before. As a result, Maryland had to

spend substantial dollars to build a replacement test covering the same content in order to ensure test security and the validity of future test results. Similar threats occur when teachers teach too closely to the test. Test preparation that targets the content too narrowly constitutes cheating. Under this circumstance, the assessment results are less likely to reflect test takers' knowledge and skills than their recall.

Another loss of test security occurs when organizations such as local newspapers seek the release of the questions and answers. As occurred in Arizona, one legal tactic is to demand disclosure of the state assessment items under the state's public records law.⁹ A state's public records law directs the disclosure of records that are owned or funded by the state. Without a clear exemption from the public records law, the state's large-scale assessment program may be compelled to release test items that could severely limit the future utility of the tests. Only a few states (i.e., Georgia, New York, and Texas) have designed their state assessments to allow for release of past test items to the public, which requires the state to build disposable assessments. These states release the assessment questions and answers to the public after the completion of the administration cycle in order to allow parents to see the test. This approach is vastly more expensive than development and repeated administration of one test or separate forms of the test over a period of years. In the latter situation, states offer limited inspection of the state assessments on a case-by-case basis, without permitting any copy or transcript of the items to be released. This approach guarantees test security and ensures that the validity of the state test is protected for future use. For most state testing agencies and their test publisher contractors, the disclosure of test items or data under public records laws is inimical to a strong accountability system and to any meaningful effort to use aggregated test results longitudinally to inform educational policy.

As the preceding examples illustrate, a pivotal role for test publishers is to safeguard educational assessments from misuse. This sentrylike role means actively ensuring that each state assessment is aligned with the curriculum and the content standards. Still, a test publisher's effort needs to be much broader than ensuring alignment. As *High Stakes* poignantly concludes, "In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes" (p. 2, executive summary). Safeguarding the state assessment also means that students should be given notice that graduation depends on passing the test; they should be provided with multiple opportunities to complete the high-stakes test successfully;

and they should be given meaningful remediation if they fail the test initially. It is crucial that test publishers change negative perceptions about the use of assessments for high-stakes purposes. Allowing such negative perceptions to persist and gain credibility can only undermine support for the use of assessment and encourage stakeholders to look for less incendiary alternatives.

Ensuring the Future of Educational Assessment

As standards-based curriculum and assessment are woven into the educational fabric, the demand for time-sensitive information will grow rapidly. The informational requirements of stakeholders seem likely to compel test publishers to expand their capabilities and look to technology to meet these and other demands. E-learning, e-testing, and web-based classrooms are a few examples of Internet-related activities that are changing the educational experience. Test publishers are in a position to oversee changes in the way educational assessments are developed, delivered, and used. Multiple-choice, open-ended response, and essay-style items can all share the assessment space with simulation tasks, video, audio, and other innovative item types. Innovative item types will provide a better understanding of how students learn, what they have learned, and how to improve their learning in the future. New learning technologies will advance efforts to improve education.

The delivery of e-testing on the Internet will almost surely compete with the paper-and-pencil test booklet for dominance of mainstream assessment.¹⁰ Web-based platforms are changing the look of adult and postsecondary education. E-learning is making lifelong learning for adults a reality. Information technology certification programs are pioneering the use of innovative item types and enhanced test security. Internet-based test preparatory and tutorial services are advancing instructional technology and influencing learning, especially as they relate to postsecondary admissions testing. Finally, the explosive growth in the use of essay-style items in state assessments for high-stakes decisions is driving the use of advanced computational linguistics techniques to score constructed writing responses. These actions already reveal the tendency of test publishers to seek technological solutions for labor-intensive, time-sensitive tasks in order to meet business and educational objectives; this trend will continue.

Although most of today's K-12 educational assessments are delivered in a paper-and-pencil medium, the signs show clearly that public school systems are migrating to online assessments. Pilot studies

of online testing are under way in the states of Oregon, Virginia, and South Dakota (Trotter, 2001a).¹¹ The speed in which technology is inserted into the educational experience will depend on its cost benefits and on funding.

Test publishers recognize that using a poorly implemented state assessment program for high-stakes decisions erodes public confidence and undermines support for education reforms. Once the NCLB is fully implemented, test publishers expect the demand for various assessments to increase by more than 50 percent (Steinberg & Henriques, 2001). The NCLB mandates testing of all students in mathematics and reading from third through eighth grade, but without any individual student consequences. Although 13 states now offer testing in grades three through eight, only nine of these states have standards-based tests (Olson, 2002). Nevertheless, with roughly 40 percent of 53 million school-age children in these six grades, the additional testing is raising some concerns about test publishers' capacities to handle all assessment needs. Many of the capacity concerns center on the timeliness and accuracy of assessment results for use in individual student decisions (Steinberg & Henriques, 2001). Technology will play a key role in addressing the substantial boost in the number of assessments administered and will be central to test publishers' efforts to provide error-free processing that is responsive to the states' time requirements for scores. Some states use the results of state tests to place students in next year's classes and to help teachers plan for next year's curriculum. In other states, testing occurs earlier in the winter or spring so that scores are received before the end of the school year. Whatever the state's needs, test publishers have always been able to meet them, and the increased role of online assessments will enhance response time and flexibility.

Use of technology to deliver large-scale assessments is not without peril. The mere shifting of the assessment from a paper-and-pencil to an online mode is grossly inadequate to stimulate permanent migration. Adoption of the online medium for assessment depends on its reconceptualization (Bennett, 1998, 1999). The key to revamping traditional assessment is to create new models of how students think and to link these models to new test designs. Such models, using innovative psychometric procedures, explain the ways students apply higher-order thinking and solve problems. Before full-fledged implementation of online testing, we must explore ways in which inequities such as unfamiliarity with or limited access to the online medium may adversely affect some students' performance. The

advantage of web-based education, and particularly online assessment, is that it can expand educational opportunities for all students. If it fails to realize such advantages, the use of the online medium for assessments will fall short of its educational and societal expectations.

With standards-based educational accountability comes a never-ending thirst for information from policymakers, educators, parents and even students. This desire for information is difficult to quench without pushing education into the twenty-first century and toward effective use of technology. Landgraf (2001) implores the educational testing community to "harness the power of technology" (p. 14) while urging the U.S. Congress to commission the development and management of Internet-delivered state assessments. The Consortium on Renewing Education (1998) boldly predicts that "new digital technology promises to change the core enterprises of schools teaching and learning profoundly influencing ways in which knowledge and information are discovered, distilled, compiled, stored, accessed, and used" (pp. 53-54). The realization of this prediction is well within reach. The near future of this realization is reason for educational test publishers to become leaders in the technological reform of education. When it comes to technology, test publishers would be wise to take a page out of the lessons learned by businesses over the years technology does not wait for those who are slow to recognize its benefits.

Conclusion

The momentum of testing is unstoppable. Test publishers will continue to play a vital role in the quest to achieve high-standards learning for all students. The role of test publishers will evolve from their present multifaceted role. The publishers' tool, the educational assessment, will provide valuable information about progress toward accountability goals and about the fit among content standards, curriculum, and instruction. Increased demand for test information will come as policymakers ratchet up the expectations for students, teachers, and school systems. Test publishers will have to devote more effort to ensuring appropriate uses of their assessments and to converting test data to better information. The appropriate uses of the assessment will also grow as test publishers introduce more advanced test designs and technical qualities to support the purposes of their assessments.

Still, the pressure of education reform will continue to bear down on educational assessment. The demands placed on assessments used for high-stakes decisions will require the next generation of tests to

possess sophisticated reporting capabilities built on innovative cognitive models and item types. When critics assert that the education reform effort is in a "testing frenzy," the discontent stems from testing that interrupts normal instructional activities and drives education policy. The key to addressing this discontent is to redouble publishers' efforts to make assessments as unobtrusive as they can be, similar to the curriculum and classroom instruction. The next generation of educational assessments will merge seamlessly into the educational experience of students.

Standards-based accountability systems raise the bar of academic expectations. At present, this is comparable to raising one side of the bar and ignoring the other side. To truly raise the bar of expectations requires delivering to students high-quality educational assessments, vastly improved teacher training and remedial support services, and a learning environment that fosters student success for all students. Education reform should point to the assessment as the gateway to educational opportunities and better life chances. As former U.S. Secretary of Education Richard Riley stated, "A quality education must be considered a key civil right for the 21st century" (OCR, 2000, p. vi). Test publishers will play a prominent role in achieving quality education for all students, whether through standardized assessments for instructional purposes or through assessments used to make high-stakes decisions.

I gratefully appreciate the critical review and insightful comments of Alan J. Thiemann and Elizabeth M. Fitzgerald. My opinions do not reflect the official position of the Association of Test Publishers.

Notes

1. For purposes of this discussion, I define a test publisher as an entity that develops or publishes education assessments using rigorous, well-accepted professional psychometric procedures. Individually, many test publishers deal with the significant issues presented in this chapter in developing their own products; collectively, they form a specific segment of the test publishing industry that must deal with such issues on a global basis.

2. The significance of this point is not lost on parents who consider education as improving their children's life chances. After grappling with low test scores and high dropout rates, the city of Carson voted to secede from the Los Angeles Unified School District. The leader of the secession movement, Carolyn Harris, said, "the future of our children and our community is at stake" ("City Voting," 2001, p. A16).

3. Although some states have developed rewards and penalties as part of their accountability system, Congress decided to eliminate this form of reinforcement from its initiative. Accordingly, the recent passage of the No Child Left Behind Act (NCLB) of 2001, a reauthorization of the Elementary and Secondary Education Act (ESEA), does not include President Bush's "proposed system of financial rewards and penalties for states based on their progress in improving student achievement" (Robelen, 2002, p. 29).

4. "World-class" refers to national educational standards that reflect a "thinking curriculum" and includes content standards that meet or exceed those of our strongest competitors (National Education Goals Panel, 1993, p. 8).

5. This discussion does not include high school graduation assessments, so-called "exit exams," because the courts have determined that special factors apply to such programs. Generally, states give students ample notice that these assessments must be passed to graduate, the tests are administered not just once but several times during a student's high school experience, and states have put in place remediation efforts to ensure that students who fail an early test have the opportunity to learn the material before being retested.

6. NCLB requires annual testing of students in mathematics and English from third grade through eighth grade. Viewed in the proper perspective, these annual tests are not considered high stakes because there are no high-stakes consequences for individual students based on the tests. They are, in fact, intended to provide parents and teachers with diagnostic information about each student, so that teachers may make changes in instruction and provide appropriate intervention or remediation based on each student's strengths and weaknesses, measured each year. Although data disaggregation by groups without any identification of individual students will occur, such programmatic evaluations are not high stakes, as that term is historically defined. See Heubert & Hauser (1999).

7. After spending more than five years drafting the *Resource Guide*, OCR finally released the document to the public in December 2000. However, it was archived by the Bush administration in January 2001. The Association of Test Publishers, who participated extensively in the drafting process, has met with the Department of Education several times since then to explore creating a public outreach program for all stakeholders using the *Resource Guide*; the reluctance of the department to implement such a program may change now that the NCLB legislation has been enacted.

8. Significantly, the OCR *Resource Guide* makes it clear that test score disparity among groups of students does not alone constitute discrimination under federal law. As then Undersecretary of OCR Norma V. Cantu stated in her "Dear Colleague" letter attached to the guide, "The guarantee under federal law is for equal opportunity, not equal results."

9. The Arizona Court of Appeals recently considered appeals by the state and the Phoenix Newspapers, Inc., seeking to review the decision of the trial court whether items from Arizona's Instrument to Measure Students (AIMS) test for graduation must be released under the state's public records law. The lower court held that certain items the state intends to use as anchor items in future tests did not have to be disclosed but

that the state had no basis to withhold disclosure of other items. Both the state and the Association of Test Publishers, as *amicus curiae*, have contended that because the state had determined to reuse the entire test form again during the period of the assessment program, all items should be protected and should not be released because that would invalidate the test and cause the state to spend additional millions of dollars building new assessments. On November 27, 2001, the Arizona Court of Appeals rendered an opinion that affirmed the decision of the trial court.

10. The proposed federally funded U.S. Open e-Learning Consortium (USOeC) would serve as a state-to-state test item exchange. All participating states would contribute one year's worth of test items to a common clearinghouse. Teachers (and parents) across the nation would have access to the item bank. They would be able to develop online assessment instruments to use as practice tests for students (Trotter, 2001b). These practice assessments would be low stakes, diagnostic, and customized. At first glance this proposed consortium is an exciting way to extend the classroom to the Internet. A potential drawback is that test publishers and test delivery organizations are not engaged at the outset in the development of the digital content (i.e., item bank) or its web-based delivery platform. It is also unclear how the proposed consortium avoids undermining the commercial activities of test publishers that are already offering online practice and diagnostic assessments to school systems.

11. The states of Georgia, Florida, and Pennsylvania are also working with test publishers to develop their online educational assessment capabilities.

References

- AERA, APA, & NCME [American Educational Research Association, American Psychological Association, and National Council on Measurement in Education]. (1999). *Standards for educational and psychological testing*. Washington, DC: APA.
- Barton, P. E. (1999). *Too much testing of the wrong kind; too little of the right kind in K-12 education*. Princeton, NJ: Educational Testing Service.
- Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Educational Testing Service.
- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, 18, 5-12.

- Carnevale, A. P., & Kimmel, E. W. (1997). *A national test: Balancing policy and technical issues*. Princeton, NJ: Educational Testing Service.
- CISA [Commission on Instructionally Supportive Assessment]. (2001a, October). *Building tests to support instruction and accountability: A guide for policymakers*. Retrieved from www.nea.org/accountability/buildingtests.html.
- CISA. (2001b, October). *Illustrative language for an RFP to build tests to support instruction and accountability*. Retrieved from www.nea.org/accountability/rfp.html
- City voting on leaving Los Angeles School District. (2001, November 2). *New York Times*, p. A16.
- Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment*. San Diego: Academic Press.
- Cizek, G. J. (2000). Pockets of resistance in the assessment revolution. *Educational Measurement: Issues and Practice*, 19, 16-23.
- Consortium on Renewing Education. (1998, November). *20/20 vision: A strategy for doubling America's academic achievement by the year 2020*. Nashville, TN: Peabody Center for Education Policy, Vanderbilt University.
- Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND.
- Hart, P. D., & Teeter, R. M. (2001). *A measured response: Americans speak on education reform*. A national public opinion survey conducted for the Educational Testing Service. Princeton, NJ: Educational Testing Service.
- Hartocollis, A. (2001, October 31). Scarsdale warned not to boycott state tests. *New York Times*, p. D2.

- ◆Heubert, J. P., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. (Report of the Committee on Appropriate Test Use, Board on Testing and Assessment, National Research Council). Washington D.C.: National Academy Press.
- ◆JCTP [Joint Committee on Testing Practices]. (2002). *Code of fair testing practices in education*. Available on *Measuring Up: An Anthology of Assessment Resources* [CD]. Also retrievable on-line: <http://aac.ncat.edu>.
- Katzman, J., & Hodas, S. (2001, September 12). A proctor for the testers? *Education Week*, 21. Retrieved from edweek.org/ew/ew_printstory.cfm?slug=02katzman.h21
- Landgraf, K. M. (2001, March 8). *Testimony before the Education Reform Subcommittee of the House Committee on Education and the Workforce on Measuring Success: Using assessments and accountability to raise student achievement*. Princeton, NJ: Educational Testing Service.
- ◆Linn (2000). Assessments and accountability. *Educational Researcher*, 29, 4-16.
- National Commission on Excellence in Education. (1983). *A nation at risk*. Washington, DC: U.S. Department of Education.
- ◆National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: transforming testing in America*. Chestnut Hill, MA: Author.
- National Education Goals Panel. (1993). *The national education goals report: Building a nation of learners, the national report* (Vol. 1). Washington, DC: Author.
- ◆OCR [Office of Civil Rights]. (2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy-makers*. Washington, DC: Author.
- Olson, L. (2002, January 9). Testing systems in most states not ESEA-ready. *Education Week*, 21, 26-27.

Phoenix Newspapers, Inc., v. Lisa Graham Keegan, the Arizona Board of Education, and the Arizona Department of Education. Case No. 1 Ca-CV-00-0284, on appeal from the Maricopa County Superior Court, Case No. CV99-021338 (April 2001).

Popham, W. J. (1999a). Educational quality: Why standardized tests don't measure up. *Educational Leadership*, 56, 8-15.

Popham, W. J. (1999b). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice*, 18, 13-17.

Popham, W. J. (2001, September 19). Standardized achievement tests: Misnamed and misleading. *Education Week*, 21, 46.

Robelen, E. W. (2002, January 9). ESEA to boost federal role in education. *Education Week*, 21, 28-31.

♦ Russell, M., & Haney, W. (2000). The gap between testing and technology in schools. *Statements*, 1, 2.

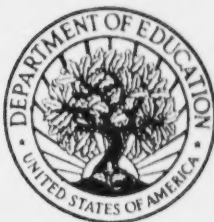
Steinberg, J., & Henriques, D. B. (2001, May 20). Right answer, wrong score: Test flaws take toll. *New York Times*, p. 1.

Trotter, A. (2001a, May 23). Testing computerized exams. *Education Week*, 20, 30-35.

Trotter, A. (2001b, November 7). Online-education consortium created for states. *Education Week*, 21, 26, 32.

Zernike, K. (2001, April 13). In high-scoring Scarsdale, a revolt against state tests. *New York Times*, p. A1.

♦ Document is included in the Anthology of Assessment Resources CD



*U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)*



NOTICE

Reproduction Basis

- ☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
- ☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").